Diário de bordo- dia 22-02-2024

Inicialmente foi apresentado o conteúdo proposto do programa da unidade Curricular, que envolve a extração do corpus com scrapping, o enriquecimento e análise de dados de várias fontes, incluindo jornais, entrevistas e imagens.

Os dados serão enriquecidos, estruturados e categorizados, e em seguida serão usados em análises linguísticas e investigações em Ciências Humanas e Sociais. Finalmente, os resultados serão visualizados usando diferentes ferramentas, em que cada aluno apresenta uma proposta de visualização.

De seguida foi apresentada a proposta dos elementos de avaliação com as respetivas percentagens, de forma a haver uma discussão que leve ao consenso de todos os alunos.

Ademais, foi também corrigido o trabalho de casa proposto na aula anterior, sendo que a resolução é a seguinte:

```python
from jjcli import *
from bs4 import BeautifulSoup as bs

ats=glob("1081-1087/Article.aspx*")
print(ats)

fo=open("saida.txt" , "w", encoding="utf-8")

def proc_article (html):
    print (len(html)) #Conta os caracteres de cada artigo
    a=bs(html) #Cria uma árvore documental
    cabecalho=""
    art= a.find("div", id="artigo") #procura

    # EXERCICIO 1- Procurar e extrair as datas, colocando-as no cabeçalho

    obter_data=art.find("span", id="ctl00_ContentPlaceHolder1_LabelInfo").text # obter o span de id ctl00_ContentPlaceHolder1_LabelInfo
    extrair_data=obter_data[:10] # obter os 10 primeiros caracteres que correspondem à data
    cabecalho += f"{extrair_data}\n" # colocar as datas no cabeçalho

    obter_slides = art.find("div", {'id':'slides'}) # ir ao div de id slides pois contém todas as imagens que queremos
    if obter_slides is not None: # caso exista div de id slides procurar pelas imagens
        for slide in obter_slides.find_all("div", {'class':'slide'}): # percorrer todos os div class slide pois contém todas as imagens que queremos
            imagem = slide.find('img') # obter a imagem
            cabecalho += f"{imagem['src']}\n" # adicionar imagem ao cabeçalho

    # EXERCICIO 2- Colocar depois do cabeçalho (---)
    cabecalho += "---\n" # colocar os ---

    for meta in a.find_all("meta"):
        p =meta.get("property")
        if p is None:
            continue
        p.replace("og" , "")
        cabecalho+=f"{p}: {meta.get('content')}\n"
       # print(p, ":" , meta.get("content"))

    # EXERCICIO 3 - Função limpeza (remover o "voltar à pagina anterior")
    for div in art.find_all("div", {'class':'voltar'}): # percorrer todos os div class voltar pois contém voltar ao inicio
        div.decompose() # apaga os voltar ao inicio

    print("==========\n", cabecalho , art.get_text(), file =fo)

for file in ats:
    with open (file, encoding="utf-8") as f:
        html=f.read()
    proc_article(html)
```

Após todos termos apresentado o nosso trabalho de casa realizado, a última aluna que apresentou, a Alexandra demonstrou as suas enormes habilidades de pesquisa extensa, mencionando os contextos das fotografias apresentadas na aula passada e fornecendo informações relevantes e bastante interessantes sobre as mesmas.

Por fim foi feita uma exploração dessas diversas imagens, que foram tiradas entre 1898 e 1901, de forma a encontrar dados ou meta dados relevantes. Foi feita uma observação dos diferentes tipos de imagens, da sua qualidade das coisas que mudaram (ex: bandeira da monarquia).

O trabalho de casa proposto para a próxima aula será então fazer uma dissecação do site da biblioteca digital, apontando os tipos de documentos que contém, bem como datas e fotografias. Após recolhidas estas informações é pedido um pequeno resumo sobre as conclusões retiradas desta observação detalhada.

Diário de bordo realizado por: Bárbara Ribeiro, pg52759